OPEN

Developing a New Version of the SF-6D Health State Classification System From the SF-36v2: SF-6Dv2

John E. Brazier, PhD, Brendan J. Mulhern, MRes,*† Jakob B. Bjorner, PhD,*‡§ Barbara Gandek, PhD,‖¶✠ Donna Rowen, PhD,* Jordi Alonso, PhD,#**†† Gemma Vilagut, PhD,#†† John E. Ware, PhD,‖¶ and on behalf of the SF-6Dv2 International Project Group*

Objective: The objective of this study was to develop the classification system for version of the SF-6D (SF-6Dv2) from the SF-36v2. SF-6Dv2 is an improved version of SF-6D, one of the most widely used generic measures of health for the calculation of quality-adjusted life years.

Study Design and Setting: A 3-step process was undertaken to generate a new classification system: (1) factor analysis to establish dimensionality; (2) Rasch analysis to understand item performance; and (3) tests of differential item function. To evaluate robustness, Rasch analyses were performed in multiple subsets of 2 large cross-sectional datasets from recently discharged hospital patients and online patient samples.

Results: On the basis of factor analysis, other psychometric evidence, cross-cultural considerations, and amenability to valuation, the 6-dimension classification used in SF-6D was maintained. SF-6Dv2 resulted in the following modifications to SF-6D: a simpler classification of physical function with clearer separation between levels; a more detailed 5-level description of role limitations; using negative wording to describe vitality; and using pain severity rather than pain interference.

Conclusions: The SF-6Dv2 classification system describes more distinct levels of health than SF-6D, changes the descriptions used for a number of dimensions and provides clearer wording for health state valuation. The second stage of the study has developed a utility value set using discrete choice methods so that the measure can be used in health technology assessment. Further work should investigate the psychometric characteristics of the new instrument.

Key Words: SF-36, SF-6D, utilities, quality adjusted life year, Rasch analysis

(*Med Care* 2020;58: 557–565)

In the economic evaluation of health interventions, the quality adjusted life year (QALY) can be used to measure outcomes. The QALY combines length and quality of life into a single figure. The quality aspect (or utility value) is anchored on a 0 (dead) to 1 (full health) scale, and can be derived from generic preference-based measures (GPBM) of health.

One such GPBM is the SF-6D (hereon SF-6Dv1)[1,2]; which was developed from version 1 of the SF-36.[3] The SF-6Dv1 describes health on 6 dimensions [physical functioning (PF), role limitations (RL), social functioning (SF), pain, mental health (MH), and vitality (VT)], 4–6 severity levels, therefore describing 18,000 health states (Fig. 1). The United Kingdom value set was developed using the standard gamble elicitation technique and ranges from 0.29 to 1.[1]

The SF-6Dv1 has become one of the most widely used GPBMs in economic evaluation.[4] Country specific value sets have been developed[5–12] and it is accepted by international reimbursement agencies.[13] SF-6Dv1 has been shown to have psychometric validity and responsiveness to change across common mental health[14,15] and physical health conditions.[16,17]

The SF-6Dv1 has not been without criticism. The severity ordering of the PF dimension (between "a lot" of limitations with moderate activities and "a little" limitation with bathing and dressing) is unclear.[1] The VT dimension is positively framed in comparison to the other dimensions. This may cause respondents

From the *School of Health and Related Research, University of Sheffield, Sheffield, UK; †Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia; ‡OptumInsight, Lincoln, RI; §Department of Public Health, University of Copenhagen, Copenhagen, Denmark; ‖John Ware Research Group, Watertown; ¶Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA; #Health Services Research Group, IMIM-Institut Hospital del Mar d´Investigacions Mèdiques; **Pompeu Fabra University (UPF); and ††CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.

✠Barbara Gandek deceased

This work was presented at the International Society for Quality of Life Research (ISOQOL) conference in Berlin, October 2014.

Supported by royalties paid for the use of version one of the SF-6D.

J.E.B. was a developer of the first version of the SF-6D, and the University of Sheffield receives royalties for the use of this instrument by commercial entities. The remaining authors declare no conflict of interest.

Correspondence to: John E. Brazier, PhD, Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK. E-mail: j.e.brazier@sheffield.ac.uk.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

ISSN: 0025-7079/20/5806-0557

confusion during valuation. The role dimension has limited sensitivity as it was based on combinations of 2 role items each with only 2 response levels. This resulted in claims of a "floor" effect, with many patients answering at the lowest severity level.[18,19] The SF-6Dv1 was also developed using SF-36v1, and there is the opportunity to revisit the classification system using the improved SF-36v2.[20,21]

Moreover, there were additional concerns raised with regard to the valuation task used to derive the SF-6Dv1 value set. On one hand, the valuation technique used for SF-6Dv1, the standard gamble is a cognitively difficult technique, and given the iterative nature of the risk trade off concerns have been raised about respondent understanding of probability and risk aversion, which may lead to higher health state values. Furthermore, the valuation task involved a 2 stage chained process with states being valued against full health and the worst state, then the worst state being valued against full health and dead, which generates higher values by doubling the impact of risk aversion. Valuation using ranking,[22] Bayesian methods[23] and discrete choice experiments, including duration (DCE_{TTO}),[13] have all produced lower values for the more severe states resulting in a wider utility range.

We therefore aimed to address these concerns by developing a new version of the SF-6Dv1 (SF-6Dv2). This includes 2 objectives: (1) to develop a "new" health state classification system from the SF-36v2 using psychometric evidence; and (2) to derive a value set that can be used in the calculation of QALYs. This paper reports on the first objective. The second objective is reported elsewhere.[24]

METHODS

Overview

We built on SF-6Dv1 and developed a new health state classification system that reflects the content of the SF-36, and produced health states amenable to valuation. To do this we used a 3-step process. Step 1 evaluated the dimensionality of the SF-36 for use in the classification system; Step 2 involved item elimination and selection; and Step 3 involved further analyses of the robustness of the Step 2 results across different data subsets. The dimension formation and item selection criteria used Factor and Rasch analyses, and other criteria such as cross-cultural relevance. Contrary to SF-6Dv1, item selection was not restricted to those also included in SF-12. This iterative methodological process was developed by a number of the authors, and has been applied widely to generate classification systems from existing profile measures.[25–27]

The SF-36

SF-36[3,28] is a measure of health that has been widely used and validated. It has 36 items across 8 dimensions: PF, role limitations due to physical health (RP), bodily pain (BP), general health (GH), VT, SF, role limitations due to emotional problems (RE), and MH. The SF-36v2 was used to develop SF-6Dv2. The SF-36v2 is an improved version of the SF-36v1. Changes included increasing the RP and RE item response levels from 2 to 5 to improve sensitivity, and simplifying the MH and VT items by reducing the levels from 6 to 5. Wording changes were also made.[20,21]

Data

Data used for this study were sourced from the 2 samples described below.

Health Outcome Data Repository Dataset[29]

Health Outcome Data Repository Dataset (HODaR) is a survey of recently discharged hospital inpatients and outpatients in the United Kingdom. The data included 49,029 full completers of the SF-36v2 between August 2002 and November 2008. The SF-36v2 was administered postally ~6 weeks after discharge, with other information linked from hospital records.

Multi Instrument Comparison Study[30]

Multi Instrument Comparison (MIC) is a survey of respondents self-reporting a range of health conditions, and a "healthy public" sample. The data used were from 5,331 respondents who fully completed the SF-36v2 online in the United Kingdom (n = 1358), Canada (n = 1335), Australia (n = 1171), and the United States (n = 1467).

The 2 samples covered a wide range of conditions and included a large proportion reporting comorbid health problems. Table 1 reports demographics and Appendix 1 (Supplemental Digital Content 1, http://links.lww.com/MLR/C7) provides detailed descriptions of the datasets.

Step 1—Dimensionality Assessment

We determined the dimensions to include in the classification system by evaluating the dimensionality of the SF-36v2 using exploratory (EFA) and confirmatory (CFA) factor analysis. This was done considering the SF-6Dv1 6-dimension structure, and we looked for evidence supporting the inclusion of more or fewer dimensions. Past work assessing SF-36 dimensionality was also considered.[31] This included research both supporting the hypothesized 8 factor structure,[32–34] and suggesting that the direction of the item response wording may impact dimensionality assessment due to response set patterns.[35,36] Work from Asia suggesting that the physical and emotional role dimensions load as one factor was also considered.[37]

The 5 GH items, and the health transition question, were not included as these are not relevant for a classification system assessing specific constructs of health. Factor analysis was conducted using Stata 15.[38]

The decision process used to select the dimensionality was also informed by face validity, conceptual coverage and cross-cultural issues. This was supported by input from the SF-6Dv2 international project team (including experts from 17 countries).

Exploratory Factor Analysis

EFA was used to identify patterns of loadings and examine dimensionality without assuming a prior structure by assessing the degree to which the item correlations can be explained by a number of factors. The number of factors to extract can be decided using eigenvalue analysis, or set by the analyst with consideration of the variance explained. We assessed the Kaiser-Meyer-Olkin measure of sample

TABLE 1. Summary of HODaR and MIC Datasets

	HODaR	MIC
Sample size	49,029	5331
Country [n (%)]		
United Kingdom	49,029 (100)	1358 (25.5)
Canada	NA	1335 (25.0)
Australia	NA	1171 (22.0)
United States	NA	1467 (27.5)
Age category (y)		
18–24	2082 (4.2)	324 (6.1)
25–34	3728 (7.6)	651 (12.2)
35–44	5560 (11.3)	726 (13.6)
45–54	7586 (15.5)	1068 (20.0)
55–64	10,568 (21.6)	1400 (26.3)
65+	19,505 (39.8)	1162 (21.8)
Female [n (%)]	24,563 (50.1)	2281 (55.9)
Education above school leaving age	NM	3467 (65.0)
Married/partner	NM	3397 (63.7)
Health status [n (%)]		
Good-excellent	28,198 (57.5)	3330 (62.5)
Fair-poor	20,831 (42.5)	2001 (37.5)
Condition [n (%)]		
Healthy population	NM	947 (17.8)
Asthma	NM	579 (10.9)
Cancer/any tumor	7202 (14.7)	577 (10.8)
Depression	NM	617 (11.6)
Diabetes	4205 (8.6)	641 (12.0)
Hearing problems	NM	601 (11.3)
Arthritis	NM	640 (12.0)
Heart disease	4129 (8.4)	706 (13.2)
COPD	5518 (11.3)	NM
Peripheral/cerebrovascular disease	2696 (5.5)	NM
Dementia	34 (0.1)	NM
Liver disease	149 (0.3)	NM
Connective tissue disease	447 (0.9)	NM
Renal disease	895 (1.8)	NM
Hemiplegia	310 (0.6)	NM
Ulcer disease	922 (1.9)	NM
HIV	46 (0.1)	NM
Report comorbidity on Charlson index	19,528 (39.8)	NM
SF-6D index		
Mean (SD)	0.66 (0.15)	0.70 (0.14)
Range	0.30–1	0.30–1

COPD indicates chronic obstructive pulmonary disease; HIV, human immunodeficiency viruses; HODaR, Health Outcomes Data Repository dataset; MIC, Multi Instrument Comparison Study dataset; NA, not applicable; NM, not measured.

adequacy, which ranges between 0 and 1, where smaller values indicate that EFA may be inappropriate.

We extracted models with 2 to 9 factors (including the model suggested by the criteria of accepting factors with eigenvalues >1) on the full samples from both datasets. To ensure reliability, models were also tested on randomly selected subsamples. For each model we assessed the adequacy of the conceptual structures for developing a classification system.

Models were developed using oblique promax rotation, which assumes dimension correlations. We used polychoric correlations due to the ordinal responses. Items loading on a dimension with a correlation of <0.4, or cross-loading between dimensions within 0.2, were identified as demonstrating poor fit, but were not excluded from the model given the aim to develop a conceptually relevant classification system. This was in line with other studies developing classification systems.[25–27]

Confirmatory Factor Analysis

CFA[39] assesses the fit of hypothesized factor structures by comparing the observed item correlation or covariance matrix with the expected matrix from the specified model. Analysis was conducted on the full sample from the HODaR and MIC datasets separately, with subsample analyses also performed. We used a range of statistics to assess model fit and guide dimensionality development. These included the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) estimates of fit, where a value of. 08 or less indicated reasonable model fit and 0.05 or less indicated good fit.[40,41] Two indices that take into account model fit and complexity, the comparative fit index (CFI)[42] and the Tucker and Lewis[43] Index (TLI), were also used. These range from 0 to 1, with values above 0.90 and 0.95 considered reasonable and excellent fit, respectively. These measures have been previously used to assess SF-12 and SF-36.[44,45]

The fit of 8 models was tested: a 7 factor model aligning with the SF-36 (PF, RP, RE, SF, BP, MH, VT for HODaR and MIC; models 1 and 2), the models produced using the selection criterion of eigenvalues >1 (models 3 and 4), and the best conceptually fitting models from EFA (models 5 and 6). CFA was also conducted on the model used for the classification system (models 7 and 8).

Step 2 and 3—Classification System Development

Rasch analysis[46] models the relationship between categorical item responses in a multi item scale and a continuous latent scale which measures an assumed underlying unidimensional construct (in this case the aspect of health measured). As Rasch models assume unidimensionality, a separate model was evaluated for each dimension. The probability of a response to each level of each item was used to assess the severity of the item against the underlying latent construct. This allowed a range of item performance indicators to be assessed. Rasch analysis was conducted using RUMM2030.[47]

We evaluated Rasch fit statistics which assess the divergence between expected and observed responses for both respondents (person-fit residuals) and items (item-fit residuals). Items with a fit residual outside the standard cutoff of ± 2.5 were considered for exclusion.

The Rasch model was also used to assess systematic differences in item response patterns for different subgroups of respondents [known as differential item functioning (DIF)]. DIF was detected using a 2-way analysis of variance assessment of the standardized residuals of the responses, where one factor was the class intervals representing severity across the latent trait scale and the other factor was the demographic subgroup.[48] We tested for DIF based on age, sex and whether a health condition was reported. There are 2 types of DIF, uniform and nonuniform. Uniform DIF occurs when a subgroup consistently differs in their responses to an item conditional on the trait estimate. Presence of uniform DIF was indicated by a significant main effect for estimates for each item across demographic groups. Nonuniform DIF occurs when the association between item responses and

group is not constant across the severity range. In the analysis this was indicated by a significant interaction effect between the subgroup and the severity range of the latent trait. A Bonferroni adjustment was applied to the significance estimates.

We additionally assessed items for inclusion in the classification system considering a range of criteria such as item severity range coverage, where a large distance from the lowest to the highest item response threshold indicates that an item provides measurement precision over a wider range.

Other criteria included avoiding complex item combinations and descriptions which may be difficult to understand and subsequently value. The face validity of the dimension wording was also evaluated.

Assessing the Robustness of the Results

To enhance robustness, Rasch and DIF analyses were conducted on 9 subsamples of combined HODaR and MIC data randomly selected for a sample of ~500 (a recommended sample size for Rasch analysis).[49] The subsamples did not significantly differ in terms of age, sex or health status. Each of the items was given an overall score (of 9) indicating the number of samples that the item performed well on. An item was considered for inclusion if it performed well on at least 5 of the 9 subsamples. However, in some cases items with lower performance remained for selection due to other criteria.

RESULTS

Step 1—Dimensionality Assessment

Exploratory Factor Analysis

Appendix 2 (Supplemental Digital Content 1, http://links.lww.com/MLR/C7) includes the EFA models with between 2 and 9 factors estimated on both datasets. The EFA models with between 2 and 6 dimensions combined a number of existing SF-36 dimensions within the factor structure, so were not suitable for developing a classification system. Models with 8 and 9 factors included redundant factors that were difficult to interpret and define.

The model including 7 factors was most in line with the original SF-36, and explained 75.1% (HODaR) and 67.8% (MIC) of the variance. The Kaiser-Meyer-Olkin of the models was high (>0.95) indicating that sample was adequate for factor analysis. Across the HODaR and MIC models, there were 5 factors consistent across the 2 datasets and 2 factors that included similar items with minor variation. Each factor is described below:

- Factor 1 (consistent): All 10 items from the PF dimension.
- Factor 2 (minor variation): Both models included the 4 RP items, and the HODaR model included the 2 SF items.
- Factor 3 (consistent): All 3 RE items.
- Factor 4 (consistent): 2 BP items.
- Factor 5 (minor variation): The HODaR model included the 5 MH items. The MIC model included the 3 negatively framed MH items.
- Factor 6 (consistent): The 4 positively framed items from the MH and VT. The 2 MH items cross-loaded with factor 5 in the HODaR data.
- Factor 7 (consistent): The 2 negatively framed VT items.

The SF dimension did not uniquely load on a single factor in either model, and failed to load at a level of 0.4 in the MIC model. The positively framed VT items loaded with the 2 positive MH items in MIC. The MH items loaded on one factor in the HODaR model, and on separate positive and negative factors in MIC.

Confirmatory Factor Analysis

Table 2 reports the CFA model fit. The worst performing models were those based on the factor structures produced using eigenvalue criteria (models 3 and 4). The models most consistent with the standard SF-36 structure (1 and 2) had CFI and TLI scores above the cutoff, but were below the models from the EFA [which would be expected given these are based on the same data samples (models 5 and 6)]. Models 1 and 2 performed better on the SRMR criteria in comparison to models 5 and 6, but lower RMSEA values were observed.

Decisions Regarding Dimensionality

PF explained the most variance and was retained. The role physical and role emotional dimensions were combined to create a single role dimension (RL) for 3 reasons. First, there is evidence from some Asian cultures that role limitations due to emotional problems is not recognized and so should not be separated out.[37] Second, CFA modeling (Table 4, models 7 and 8) suggested that the correlation coefficients for the combined role dimension were acceptable. Third, including separate role physical and role emotional dimensions could complicate the valuation process. Although social limitations were correlated with RLs, the importance of general social activities with friends and family, and the likely impact of health interventions on social aspects, led to retaining this as an individual dimension assessing general SF. The 2 pain items formed a strong factor in both samples. These first 4 dimensions were negatively worded, and it was decided to ensure MH and VT were consistent with this. There was evidence from MIC that the positively worded items for MH overlap with VT. For MH the negatively worded items were part of the MH factor in HoDAR and formed a

TABLE 2. Confirmatory Factor Analysis Model Fit Statistics

	Model Number							
	1	2	3	4	5	6	7	8
RMSEA	0.084	0.088	0.152	0.147	0.080	0.079	0.085	0.078
SRMR	0.041	0.043	0.068	0.077	0.048	0.048	0.036	0.043
CFI	0.917	0.901	0.816	0.806	0.926	0.923	0.934	0.932
TLI	0.905	0.886	0.792	0.781	0.914	0.911	0.923	0.920
N	49,029	5331	49,029	5331	49,029	5331	49,029	5331

Models 1 and 2: 7 factor models aligning with the SF-36 (PF, RP, RE, SF, BP, MH, VT, excluding GH) for HODaR and MIC, respectively.

Models 3 and 4: 3 factor models produced using the factor selection criterion of eigenvalues >1 for the HODaR and MIC.

Models 5 and 6: Best conceptually fitting models identified from the EFA for the HODaR and MIC data with 7 dimensions.

Models 7 and 8: The final 6 factor model used for health state classification system run on the HODaR and MIC data.

BP indicates bodily pain; CFI, Comparative Fit Index; GH, general health; HODaR, Health Outcome Data Repository Dataset; MH, mental health; MIC, Multi Instrument Comparison; PF, physical functioning; RMSEA, root mean square error of approximation; SF, social functioning; SRMR, standardized root mean square residual; TLI, Tucker-Lewis Index; VT, vitality.

single factor in MIC. The negatively framed items were retained as a single MH dimension. The negatively framed VT items formed their own factor in both datasets, in line with earlier work,[36] so were used as a simplified description of VT.

This 6 dimensional structure was in line with that used for SF-6Dv1. Table 2 reports the CFA model fit results based on the selected factor structure (models 7 and 8). These models had acceptable CFI and TLI scores. Model 7 had the lowest overall SRMR and model 8 had the lowest RMSEA. Table 3 reports the factor loadings from the CFA models for the final 6 factor model, and shows that all coefficients were above the minimum level. This provides support for retaining the 6 dimensional structure as the basis for the classification system.

TABLE 3. Confirmatory Factor Analysis Loadings for 6 Factor Model

	CFA Loading*	
Dimension/Item	**HODAR**	**MIC**
Physical functioning		
Q3. Limited vigorous activities	0.665	0.677
Q4. Limited moderate activities	0.845	0.837
Q5. Limited lifting or carrying groceries	0.832	0.804
Q6. Limited climbing several flights of stairs	0.860	0.842
Q7. Limited climbing one flight of stairs	0.854	0.853
Q8. Limited bending kneeling or stooping	0.794	0.765
Q9. Limited walking >1 mile	0.876	0.868
Q10. Limited walking several hundred yards	0.896	0.878
Q11. Limited walking 100 yards	0.834	0.813
Q12. Limited bathing or dressing	0.711	0.609
Role physical		
Q13. Cut down time spent on work/other activities	0.923	0.888
Q14. Accomplished less than you would like	0.945	0.908
Q15. Limited in the kind of work/other activities	0.952	0.937
Q16. Difficulty performing work or other activities	0.956	0.944
Role emotional		
Q17. Cut down time spent on work/other activities	0.637	0.667
Q18. Accomplished less than you would like	0.644	0.643
Q19. Did work or other activities less carefully	0.631	0.645
Social functioning		
Q20. Physical/emotional health interfere with social activities (extent)	0.903	0.859
Q32. Physical/emotional health interfere social activities (frequency)	0.903	0.880
Pain		
Q21. Severity of bodily pain	0.784	0.831
Q22. Extent pain interfered with normal work	0.981	0.959
Mental health—negatively worded items		
Q24. Very nervous	0.723	0.719
Q25. Down in the dumps	0.878	0.871
Q28. Downhearted and depressed	0.874	0.895
Vitality—negatively worded items		
Q29. Feel worn out	0.890	0.908
Q31. Feel tired	0.835	0.834

*The loading values in the table are taken from the CFA model testing the final 6 factor structure used to select items for the classification system.

CFA indicates confirmatory factor analysis; HODaR, Health Outcome Data Repository Dataset; MIC, Multi Instrument Comparison.

Steps 2 and 3—Item Selection and Health State Classification System Development

The process for selecting items for each dimension is outlined below. Table 4 displays the results including the overall score for the number of times the item performed well of 9 subsamples, and scores regarding the times DIF or poor fit was exhibited. Appendix 3 (Supplemental Digital Content 1, http://links.lww.com/MLR/C7) includes an item-by-item summary of their performance, including the stage at which the item was excluded, and which were retained.

Physical Functioning

Of the 10 items from the SF-36 PF scale, there was evidence of misfit for 4 items. Two items displayed DIF by sex.

Of the remaining 4 items, the vigorous activity and bathing and dressing limitation items were selected to ensure the dimension was sensitive to the less severe (mean latent scale coverage 1.63 to 3.78) and more severe (coverage −3.71 to −1.42) range, respectively. As these items did not overlap in values on the latent scale, the item assessing moderate activity limitations was chosen to cover the mid-range of the severity scale (mean range −0.88 to 1.64). This item did demonstrate misfit in a number of the analyses. However, retaining this item in the classification system increased sensitivity across the severity range.

Role Limitations

Across the role physical and role emotional dimensions, all 7 items had good Rasch statistics on the majority of analyses and covered a similar severity range. A key criterion was face validity, and the physical and emotional role dimensions were combined to tackle cross-cultural factors linked to evidence that RL due to emotional problems is not recognized in some Asian countries. Therefore, it was decided to use items assessing the same concept across both physical and emotional RLs. This resulted in the item "Accomplished less than you would like" being used for both but with attribution to either physical health or emotional problems.

Social Functioning

Both items performed well, and covered a similar severity range, with the average position centrally located on the logit scale (0.01 and −0.01). Because of this, and for consistency with SF-6Dv1, it was decided to retain the social activity limitation frequency item.

Pain

There was more evidence of misfit for pain interference (7/9 samples) compared to pain severity (4/9 samples). The pain severity item covered a wider range, particularly at milder severity. Furthermore, it avoided reference to interference with work and correlated less with the PF, RP, and RE dimensions. For these reasons pain severity was selected.

Mental Health

The item "down in the dumps" displayed item misfit. The negatively worded items for assessing frequency of being "very nervous" and "down and depressed" remained for selection following the Rasch analysis and they were combined as one dimension within the classification system.

TABLE 4. Rasch Analysis, and Summary of Chosen Dimension Structure

Dimension/Item	Overall Score*	Mean Item Level Performance				DIF‖	Item Misfit¶	Selected
		Range†	Location‡	Fit Residual§	*P*			
Physical functioning								
Q3. Limited vigorous activities	8/9	1.63–3.78	2.71	0.23	0.28	Sex (1 sample)	No misfit	√
Q4. Limited moderate activities	3/9	−0.88 to 1.64	0.38	−1.31	0.17	No DIF evidence	Misfit (6 samples)	√
Q5. Limited lifting or carrying groceries	2/9	−1.78 to 1.29	−0.25	−1.05	0.10	Sex (6 samples)	Misfit (1 sample)	X
Q6. Limited climbing several flights of stairs	1/9	−0.50 to 2.30	0.90	−2.06	0.17	Sex (3 samples)	Misfit (5 samples)	X
Q7. Limited climbing one flight of stairs	3/9	−2.46 to 0.33	−1.07	−1.92	0.10	No DIF evidence	Misfit (6 samples)	X
Q8. Limited bending kneeling or stooping	9/9	−1.02 to 1.95	0.47	−0.04	0.15	No DIF evidence	No misfit	X
Q9. Limited walking > 1 mile	2/9	0.087–1.43	0.76	−2.34	0.04	Age (1 sample)	Misfit (6 samples)	X
Q10. Limited walking several hundred yards	0/9	NA#	NA	NA	NA	No DIF evidence	Misfit (9 samples)	X
Q11. Limited walking 100 yards	6/9	−2.32 to −0.55	−1.44	−0.99	0.18	No DIF evidence	Misfit (3 samples)	X
Q12. Limited bathing or dressing	8/9	−3.71 to −1.42	−2.56	0.19	0.22	Age (1 sample)	No misfit	√
Role physical								
Q13. Cut down time spent on work/other activities	8/9	−3.34 to 2.46	−0.53	1.38	0.28	No DIF evidence	Misfit (1 sample)	X
Q14. Accomplished less than you would like	9/9	−3.16 to 3.87	0.31	0.28	0.30	No DIF evidence	No misfit	√
Q15. Limited in the kind of work/other activities	7/9	−2.84 to 3.11	0.10	−1.44	0.34	No DIF evidence	Misfit (2 samples)	X
Q16. Difficulty performing work or other activities	9/9	−3.04 to 3.07	0.08	−0.96	0.39	No DIF evidence	No misfit	X
Role emotional								
Q17. Cut down time spent on work/other activities	9/9	−3.44 to 2.15	−0.20	−0.63	0.41	No DIF evidence	No misfit	X
Q18. Accomplished less than you would like	7/9	−3.33 to 4.25	0.59	−0.90	0.09	Age (1 sample)	Misfit (1 sample)	√
Q19. Did work or other activities less carefully	9/9	−3.58 to 2.75	0.34	1.07	0.31	No DIF evidence	No misfit	X
Social functioning								
Q20. Physical/emotional health interfere with social activities (extent)	9/9	−2.88 to 2.75	0.01	0.48	0.30	No DIF evidence	No misfit	X
Q32. Physical/emotional health interfere social activities (frequency)	9/9	−2.81 to 2.51	−0.01	0.61	0.23	No DIF evidence	No misfit	√
Pain								
Q21. Severity of bodily pain	5/9	−4.74 to 4.70	0.29	−0.51	0.13	No DIF evidence	Misfit (4 samples)	√
Q22. Extent pain interfered with normal work	1/9	−2.20 to 2.33	−0.04	0.58	0.19	Sex (1 sample)	Misfit (7 samples)	X
Mental health—negatively worded items								
Q24. Very nervous	6/9	−2.25 to 2.17	−0.16	1.05	0.24	No DIF evidence	Misfit (3 samples)	√
Q25. Down in the dumps	3/9	−3.11 to 2.18	−0.30	−0.36	0.08	No DIF evidence	Misfit (6 samples)	X
Q28. Downhearted and depressed	9/9	−1.78 to 2.56	0.21	0.19	0.22	No DIF evidence	No misfit	√
Vitality—negatively worded items								
Q29. Feel worn out	8/9	−3.56 to 1.50	−0.88	0.44	0.12	No DIF evidence	Misfit (1 sample)	√
Q31. Feel tired	7/9	−2.75 to 3.33	0.08	−0.07	0.34	Sex (1 sample)	Misfit (1 sample)	X

*Number of the 9 Rasch analyses the item was valid for across each of the criteria (high scores better).
†Range on the latent scale.
‡Mean location parameter.
§Fit item-fit residual.
‖DIF (sample number is the number of sample the item exhibited DIF on across the 9 tested).
¶Number of times that particular item did not fit the Rasch model across the 9 samples and so was removed from the Rasch model.
#Results not available as the items was excluded as did not perform well on any of the 9 analyses.
X indicates item excluded from classification system; √, item included in classification system; DIF, differential item function; NA, not applicable.

SF-6Dv1	SF-6Dv2
PHYSICAL FUNCTIONING	
Your health does not limit you in vigorous activities	Limited in vigorous activities not at all
Your health limits you a little in vigorous activities	Limited in vigorous activities a little
Your health limits you a little in moderate activities	Limited in moderate activities a little
Your health limits you a lot in moderate activities	Limited in moderate activities a lot
Your health limits you a little in bathing and dressing	Limited in bathing and dressing a lot
Your health limits you a lot in bathing and dressing	
ROLE LIMITATIONS	
No problems with your work or other daily activities as a result of your physical health or any emotional problems	Accomplish less than you would like none of the time
Limited in the kind of work or other activities as a result of your physical health	Accomplish less than you would like a little of the time
Accomplish less than you would like as a result of emotional problems	Accomplish less than you would like some of the time
Limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems	Accomplish less than you would like most of the time
	Accomplish less than you would like all of the time
SOCIAL FUNCTIONING	
Your health limits your social activities none of the time	Social activities are limited none of the time
Your health limits your social activities a little of the time	Social activities are limited a little of the time
Your health limits your social activities some of the time	Social activities are limited some of the time
Your health limits your social activities most of the time	Social activities are limited most of the time
Your health limits your social activities all of the time	Social activities are limited all of the time
PAIN	
You have no pain	No pain
You have pain but it does not interfere with your normal work	Very mild pain
You have pain that interferes with your normal work a little bit	Mild pain
You have pain that interferes with your normal work moderately	Moderate pain
You have pain that interferes with your normal work quite a bit	Severe pain
You have pain that interferes with your normal work extremely	Very severe pain
MENTAL HEALTH	
You feel tense or downhearted and low none of the time	Depressed or very nervous none of the time
You feel tense or downhearted and low a little of the time	Depressed or very nervous a little of the time
You feel tense or downhearted and low some of the time	Depressed or very nervous some of the time
You feel tense or downhearted and low most of the time	Depressed or very nervous most of the time
You feel tense or downhearted and low all of the time	Depressed or very nervous all of the time
VITALITY	
You have a lot of energy all of the time	Worn out none of the time
You have a lot of energy most of the time	Worn out a little of the time
You have a lot of energy some of the time	Worn out some of the time
You have a lot of energy a little of the time	Worn out most of the time
You have a lot of energy none of the time	Worn out all of the time

FIGURE 1. SF-6D health state classification system comparison.

Vitality

Both the negatively framed VT items performed well. The item assessing being "worn out" was selected over the more general and less severe "tired" item.

SF-6Dv2 Classification System

The SF-6Dv2 classification system is displayed in Figure 1, with SF-6Dv1 for comparison. The figure shows the differences in the items selected, and the simplification of the dimension level wording to support valuation. This was done through an iterative process, led by authors J.E.B., B.J.M., and D.R., of adapting the item content and item level wording into single sentences reflecting the original meaning of the item. The process involved review and revision of potential changes by the other authors and the international SF-6Dv2 team. Appendix 4 (Supplemental Digital Content 1, http://links.lww.com/MLR/C7) demonstrates how the selected SF-36v2 items were converted into the SF-6Dv2 classification system. In constructing the classification, the team was aware that the initial valuation will use DCE_{TTO} and this requires respondents to compare states varying in content. The simplification in wording was achieved by providing instructions at the beginning of the task that the descriptions are because of health.

The key changes in comparison to SF-6Dv1, are:

- PF uses the same items but reduces the levels from 6 to 5 to avoid ambiguity in level ordering by removing "Limited a little in bathing and dressing".
- RL descriptions are simplified by using the same item to represent the 2 constituent dimensions. The increase to 5 levels takes advantage of the greater sensitivity of SF-36v2.
- SF is the same as for SF-6Dv1, with simplified level descriptions.
- Pain has changed from interference to severity to increase sensitivity to changes in pain intensity.
- MH includes both depression and anxiety focused items in line with SF-6Dv1.
- VT has changed from positively worded "energy" to negatively worded "worn out."

CONCLUSIONS

In this study we have developed a new version of the SF-6D classification system (SF-6Dv2) informed by psychometric results and considering previous work testing

the limitations of SF-6Dv1. This process means that the SF-6Dv2 should overcome many of the published criticisms of SF-6Dv1 but maintain similarities. This is because we have used the same dimension structure and retained much of the descriptive content, whilst simplifying wording to support valuation using DCE_{TTO}. We also involved international experts to ensure cultural issues were considered. Some exploratory comparisons of the SF-6Dv1 and SF-6Dv2 using the preferred value sets are reported elsewhere.[24]

There are improvements between versions to standardize the direction of the wording (VT), simplify the wording (RL), remove inconsistencies (PF) and move towards the measurement of pain severity (PA). Standardizing the direction of the wording has been shown to increase the psychometric validity of tests.[50] Item level changes will impact on the ability to measure change in health status over time.

This study has a number of limitations and areas for future work. First, we used datasets that were restricted to westernized majority English speaking countries. To overcome this the international expert group included researchers involved in the original development and translation of SF-36. This was influential in the decision to combine the role dimensions. DIF analysis based on a wider range of demographics (such as language, education and country of residence) was not possible as we were restricted to the data collected in the HODaR and MIC studies. In contrast to the development of SF-6Dv1, we did not restrict the item selection to those included in SF-12. SF-6Dv2 values will be estimated from SF-12 using mapping algorithms in line with methods used elsewhere.[51]

Further work is required to psychometrically test SF-6Dv2 in comparison to SF-6Dv1, particularly the responsiveness of items to change. This will allow for further understanding of the new classification system to support widespread use of the instrument. A key comparison is with the EQ-5D-5L.[52,53] There are important differences between the SF-6Dv2 and EQ-5D-5L descriptive systems that will result in different psychometric indicators. Comparisons in patient datasets will help understand the relationship between the measures, and inform the use of each in health technology assessment.

In conclusion, we have developed a simplified version of the SF-6D classification system considering many of the criticisms of SF-6Dv1. We have used updated psychometric methods that allow insight into the choices made during the development of SF-6Dv1. This led to the same dimension structure, but with improvements. The classification system will be valued internationally using valuation methods based on DCE. The United Kingdom valuation is reported in a companion paper,[24] and subsequent international value sets can be tested for use in the estimation of QALYs for the economic evaluation of health interventions.

ACKNOWLEDGMENTS

The authors would like to acknowledge the passing of their dear colleague Barb Gandek during the final writing up of this research. She will be greatly missed by all. The authors would like to acknowledge the input of the remaining SF-6Dv2 international project team which includes: Nick Bansback, Beate Bestmann, Luciane Cruz, Rajabali Daroudi, Lara Ferreira, Pedro Ferreira, Shunichi Fukuhura, Lewis Kazis, Thomas Kohlmann, Maria Knoph Kvamme, Cindy Lam, Clara Mukuria, Richard Norman, Jan Abel Olsen, Julie Ratcliffe, Antonio Rosello, Akbari Sari, Rick Sawatsky, Elly Stolk, Dong Suh, Gemma Vilagut, David Whitehurst, Carlos Wong, Jing Wu, Yosuke Yamamoto. They would also like to thank the HODaR and MIC project teams for use of their data in this study.

REFERENCES

1. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ*. 2002;21:271–292.
2. Brazier JE, Roberts J. Estimating a preference-based index from the SF-12. *Med Care*. 2004;42:851–859.
3. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection. *Med Care*. 1992; 30:473–483.
4. Wisløff T, Hagen G, Hamidi V, et al. Estimating QALY gains in applied studies: a review of cost utility analyses published in 2010. *Pharmacoeconomics*. 2014;32:367–375.
5. Abellán Perpiñán JM, Sánchez Martínez FI, Martínez Pérez JE, et al. Lowering the 'floor' of the SF-6D scoring algorithm using a lottery equivalent method. *Health Econ*. 2012;21:1271–1285.
6. Brazier J, Fukuhara S, Roberts J, et al. Estimating a preference-based index from the Japanese SF-36. *J Clin Epidemiol*. 2009;62:1323–1331.
7. Cruz L, Camey S, Hoffmann JF, et al. Estimating the SF-6D value set for a population-based sample of Brazilians. *Value Health*. 2011;14: S108–S114.
8. Ferreira LN, Ferreira PL, Pereira LN, et al. A Portuguese value set for the SF-6D. *Value Health*. 2010;13:624–630.
9. Jonker MF, Donkers B, de Bekker-Grob EW, et al. Advocating a paradigm shift in health-state valuations: the estimation of time-preference corrected QALY Tariffs. *Value Health*. 2018;21:993–1001.
10. Lam CL, Brazier J, McGhee SM. Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a Chinese population. *Value Health*. 2008;11:295–303.
11. McGhee SM, Brazier J, Lam CL, et al. Quality-adjusted life years: population-specific measurement of the quality component. *Hong Kong Med J*. 2011;17:17–21.
12. Norman R, Viney R, Brazier J, et al. Valuing SF-6D health states using a discrete choice experiment. *Med Decis Making*. 2014;34:773–786.
13. International Society for Pharmacoeconomics & Outcomes Research. Pharmacoeconomic guidelines around the world [Online]. 2019 Available at: https://tools.ispor.org/peguidelines/. Accessed May 5, 2019.
14. Brazier JE, Connell J, Papaioannou D, et al. Validating generic preference-based measures of health in mental health populations and estimating mapping functions for widely used specific measures. *Health Technol Assess*. 2014;18:1–188.
15. Mulhern B, Mukuria C, Barkham M, et al. Using preference-based measures in mental health conditions: the psychometric validity of the EQ-5D and SF-6D. *Br J Psychiatry*. 2014;205:236–243.
16. Longworth L, Yang Y, Young T, et al. Use of generic and condition specific measures of Health Related Quality of Life in NICE decision making. *Health Technol Assess*. 2014;18:1–224.
17. Brazier JE, Tsuchiya A, Roberts J, et al. A comparison of the EQ-5D and the SF-6D across seven patient groups. *Health Econ*. 2004;13: 873–884.
18. Ferreira PL, Ferreira LN, Pereira LN. How consistent are health utility values? *Qual Life Res*. 2008;17:1031–1042.
19. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Econ*. 2003;12:1061–1067.
20. Ware JE, Kosinski M, Dewey JE. *How to Score Version Two of the SF-36 Health Survey*. Lincoln, RI: QualityMetric Incorporated; 2000.
21. Ware JE, Kosinski M, Bjorner JB, et al. *User's Manual for the SF 36v2® Health Survey*, 2nd ed. Lincoln, RI: QualityMetric Incorporated; 2007.
22. McCabe C, Brazier J, Gilks P, et al. Using rank data to estimate health state utility models. *J Health Econ*. 2006;25:418–431.

23. Kharroubi SA, Brazier JE, Roberts J, et al. Modelling SF-6D health state preference data using a nonparametric Bayesian method. *J Health Econ*. 2007;26:597–612.
24. Mulhern B, Norman R, Bansback N, et al. Valuing SF-6Dv2 in the UK using a discrete choice experiment with duration. *Med Care*. 2020. Doi: 10.1097/MLR.0000000000001324.
25. Brazier JE, Rowen D, Mavranezouli I, et al. Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient based measures of outcome). *Health Technol Assess*. 2012;16:1–114.
26. Mulhern B, Rowen D, Jacoby A, et al. The development of a QALY measure for epilepsy: NEWQOL-6D. *Epilepsy Behav*. 2012;24:36–43.
27. Rowen D, Brazier J, Young T, et al. Deriving a preference based measure for cancer using the EORTC QLQ-C30. *Value Health*. 2011;14:721–731.
28. Frendl DM, Ware JE. Patient-reported functional health and well-being outcomes with drug therapy: a systematic review of randomized trials using the SF-36 health survey. *Med Care*. 2014;52:439–445.
29. Currie CJ, McEwan P, Peters JR, et al. The routine collation of health outcomes data from hospital treated subjects in the Health Outcomes Data Repository (HODaR): descriptive analysis from the first 20,000 subjects. *Value Health*. 2005;8:581–590.
30. Richardson J, Iezzi A, Maxwell A. *Cross-national Comparison of Twelve Quality of Life Instruments: MIC Paper 1 Background, Questions, Instruments*. Melbourne, Australian: Centre for Health Economics, Monash University; 2012.
31. De Vet HC, Ader HJ, Terwee CB, et al. Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. *Qual Life Res*. 2005;14:1203–1218.
32. McHorney CA, Ware JE, Lu JF, et al. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care*. 1994;32:40–66.
33. Keller SD, Ware JE, Bentler PM, et al. Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol*. 1998;51:1179–1188.
34. Ware JE, Kosinski M, Gandek B, et al. The factor structure of the SF-36 Health Survey in 10 countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol*. 1998;51:1159–1165.
35. Bjorner JB, Ware JE, Kosinski M. The potential synergy between cognitive models and modern psychometric models. *Qual Life Res*. 2003;12:261–274.
36. Deng N, Rick G, Ware JE. Energy, fatigue, or both? A bifactor modeling approach to the conceptualization and measurement of vitality. *Qual Life Res*. 2015;24:81–93.
37. Suzukamoa Y, Fukuhara S, Green J, et al. Validation testing of a three-component model of Short Form-36 scores. *J Clin Epidemiol*. 2011;64:301–308.
38. StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LP; 2011.
39. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assess*. 1995;7:286–299.
40. Hu LT, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Modeling*. 1999;6:1–55.
41. Kline RB. *Principles and Practice of Structural Equation Modeling*, 2nd ed. New York, NY: Guilford; 2005.
42. Bentler PM. Comparative fit indexes in structural models. *Psychol Bull*. 1990;107:238–246.
43. Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*. 1973;38:1–10.
44. Okonkwo O, Roth D, Pulley L, et al. Confirmatory factor analysis of the validity of the SF-12 for persons with and without a history of stroke. *Qual Life Res*. 2010;19:1323–1331.
45. Su CT, Ng HS, Yang AL, et al. Psychometric Evaluation of the Short Form 36 Health Survey (SF-36) and the World Health Organization Quality of Life Scale Brief Version (WHOQOL-BREF) for Patients With Schizophrenia. *Psychol Assess*. 2014;26:980–989.
46. Rasch G. *Probabilistic Model for Some Intelligence and Achievement Tests*. Copenhagen, Denmark: Danish Institute for Educational Research; 1960.
47. Andrich D, Lyne A, Sheridan B, et al. *RUMM 2030*. Perth: RUMM Laboratory; 2010.
48. Hagquist C, Andrich D. Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health Qual Life Outcomes*. 2017;15:181.
49. Linacre JM. Sample size and item calibration stability. *Rasch Meas Trans*. 1994;7:328.
50. Burkner PC, Schulte N, Holling H. On the statistical and practical limitations of Thurstonian IRT models. *Educ Psychol Meas*. 2019;79:827–854.
51. Wailoo AJ, Hernandez-Alava M, Manca A, et al. Mapping to estimate health-state utility from non–preference-based outcome measures: an ISPOR Good Practices for Outcomes Research Task Force Report. *Value Health*. 2017;20:18–27.
52. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20:1727–1736.
53. Devlin N, Shah K, Feng Y, et al. Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England. *Health Econ*. 2018;27:7–22.

